Predicting the diffusion of brand stories on social networks

Thi Bich Ngoc Hoang [1,2] and Josiane Mothe[1]

[1] IRIT, CNRS UMR5505, UT2J, Universit de Toulouse, France
[2]University of Economics, The University of Danang, Vietnam

{thi-bich-ngoc.hoang,josiane.mothe}@irit.fr

Abstract. The emergence and growing of social media allows one consumer to communicate with thousands or millions other consumers. The consumer-generated stories about a brand or a product can be widely propagated and as a consequence can have a big impact on the marketplace and indirectly affect the success of the brand. Therefore, modeling the information diffusion in social media is crucial for business managers in order to both understand the information propagation and to better control it. Our research aims at predicting whether a tweet about a brand is going to be diffused and the level of the diffusion. We apply several machine learning classifiers using user-based, time-based and content-based features associated to tweets and developed several new features some content-based. We show that our method significantly improves F-measure by about 4% compared to the state of art. We also show that the numbers of a user's followers, number of communities that this user belongs to, and number of likes that a user has made on his time line are the most important features for the predictive model.

Keywords: Information retrieval, Information diffusion, Tweets analysis, Predictive model, Using machine learning, Online Marketing

1   Introduction

The popularity of on line social networks has rapidly increased over the past few years. social networks allows users to connect with new people, share opinions and information. While many studies focused on event-related applications, recently, a few studies focused on social networks in marketing and show that using social networks opens several new opportunities for businesses to market their product [1]. According to Mangole *et al.* since the social media allows one person to communicate with other thousands or millions people about products or brands, the impact of customer-to-customer communications has increased and managers should start brand stories to be followed by customers or contribute to existing discussions in a way that serves the business and performance goal [2]. Similarly, Gensler *et al.* [3] consider that social media significantly affects the brand management because of its dynamic, ubiquitous and regular interaction. Consumers are becoming pivotal authors of brand stories. Such stories can

create advertisements that are more effective than usual advertisements created by company-generated stories.

One of the advantages of social networks is that it can help businesses reaching their potential customers easily. According to Twitter Stats for Businesses [1], 65.8% of U.S. companies are now using Twitter for marketing purposes. As in the same source, 47% of people who follow a brand on Twitter are more likely to visit that company's website. During discussions among consumers on social networks, stories about products or brands are formed and spread thanks to the retweet functionality. By repeating the message, all user's followers are able to read the message, thus helping the message/brand story to be broad casted and reach a large amount of people.

Our work aims at helping business managers to understand and predict the diffusion of a given post in social networks as well as which features make a message popular. From that, they can join a discussion or create one and contribute in order to be consistent with businesses' mission and goal.

In this paper, we study two related research questions: (1) Is it possible to predict whether a tweet about a brand story is going to be spread i.e. re-tweeted? and (2) Can the level of diffusion be modeled and thus can we predict the level of diffusion of a new tweet that is advertising a specific product?

We answer these research questions by considering a model that makes use of various tweet features, we train on a subset of tweets and test on new tweets using different types of machine learning algorithms. While some features come from previous work in the domain of tweet diffusion [4], we also introduce new features and evaluate the added value of these new features both to predict whether a tweet is going to be retweeted and to predict the level of the propagation. We show that, we significantly improves by about 4% F-measure compared to the state of art methods for predicting retweetability of a tweet when evaluating our model on tweet collections about a brand stories generated by consumers and by the owner of the brand.

The rest of the paper is organized as follows: Section 2 presents the related work. Section 3 describes the features we used, giving a specific focus on the new features that we developed, as well as the predictive model. Section 4 presents the evaluation framework and data used. Section 5 presents the results. Finally, Section 6 is the discussions and conclusion.

2 Related Work

Suh *et al.* studied a number of features that may correlate with the retweetability of a given tweet. They considered the content and context features represented for tweets from a large-scale dataset of 74 million tweets. They showed that the numbers of followers, followees, and ages of the account have a strong relationship with the retweet number while the total tweets that a user wrote in the past has little or no relationship with the average number of daily tweets or with the

---

[1] https://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/

retweet rate. They also showed that the presence of hashtag or URL in a tweet does not highly impact on the number of retweets: 20.8% of retweets only contain hashtags while 28.4% of retweets contain URL [4]. In our work, we consider all the features proposed by Suh *et al.*. We also add several new features including user-based, time-based, and content-based features.

Hu *et al.* proposed a method for predicting the short-term popularity of viral topics utilizing a time series feature space. They used data of historical popularity of a given topic and considered three types of features: previous-popularity-based, user-comment-based and network-structure-based. They showed that the popularity is relatively dynamic and changeable for burst topics and historical popularity can still have an impact on later popularity for non-burst topics [5].

Kwa *et al.* examined the relationship between the number of followers of a user and the popularity of his or her message on social networks. They showed that people only retweets from a small number of people and only a subset of a users followers actually retweet. In addition, users with less than 1,000 followers tend to have the same average number of retweets for their posts [6]. Similarly, Remy *et al.* studied the correlation between the number of users' followers and the capacity to spread their messages. They showed that the impact of users with a lot of followers is not statistically greater than users with a few followers [7]. In our model, we also take into account the relationship between the number of followers of a user and the retweetability of his or her tweets.

Hong *et al.* addressed the problem of predicting the future retweet number of a given tweet. They formulated the task into binary classification and multi-class classification. For multi-class classification, the authors suggested 4 classes: class-0 (not retweet), class-1 (retweet number less than 100), class-2 (retweet number less than 10,000), and class-3 (retweet number greater than 10,000). They used logistic regression as a classifier considering the content of the message, structural properties of the users' social graph and meta-data of users. They achieved good accuracy only for the smallest and largest categories (class-0 and class-3) [8], but very low accuracy for the other two categories (0.15 for class-1 and 0.43 for class-3). Besides, the authors did not describe the features they used explicitly. Our idea of classifying tweets into classes is similar to Hongs'. In the evaluation section of our paper (Section 5), we show that we can improve the recall, F-measure for the binary classification by using random forest as the machine learning algorithm and several new features we introduced. In multi-class classification, we also improve the F-measure for class-1 and class-2 which are supposed to be more challenging classes since the majority of tweets are in these two classes.

Other works related to information diffusion on social networks are [9–11]. Zhang *et al.* addressed the problem of how the friends of a user impact on his or her behaviors. They found that the fraction of active users (retweeted a message) with two active neighbors (followees who have retweeted the same message) is about double compared to the fraction of active users with only one active neighbors [10]. In a similar work [11], Yang found that almost 25.5% of the tweets posted by users are actually retweeted from their friends' posts. We

did not consider the influence of followers' retweeting behavior on friends in our work since the datasets we use do not contain information of users' followers; this could be an interesting feature for our future work.

3   Predicting information diffusion: features and model

3.1   Tweet representation

We hypothesize that both the tweet content and the user who wrote it have an impact on tweet diffusion. In our model, tweets are represented by user-based, time-based and content-based features. There are a total of 32 features which are shortly described in the Table 1.

We reuse all the features Suh *et al.* [4] suggested and marked them with a[+] in Table 1. We also define some new features which is one contribution of this paper.

Table 1. Features used to predict retweet rate of a given tweet. Features with a[+] are Suh *et al.*' [4] while the other features are new features we developed.

	Features	Description	Data Type
User-based	1. Total_of_tweets[+]	Total of past tweets that the user has posted in the timeline	#Numeric
	2. No_of_followers[+]	Number of people who follow the user	#Numeric
	3. No_of_followees[+]	Number of people the user follows	#Numeric
	4. Age_of_account[+]	Number of days since the user account has been created	#Numeric
	5. No_of_favourite[+]	Number of tweets the user has liked in the timeline	#Numeric
	6. No_groups_user_belongs	Number of groups that the user belongs to	#Numeric
	7. Aver_favou_per_day	Average of likes that the user has made per day	#Numeric
	8. Aver_tweets_per_day	Average of tweets that the user has posted per day	#Numeric
	9. User_is_well_known	The user's name has been introduced on WikiPedia	#Numeric
	10. User_is_verified	The user account is verified	#Numeric
	11. User_name_len	The length of the user's name	#Numeric
Time-based	12. Is_post_at_hol	The tweet is created on public holiday	Boolean
	13. Is_posted_at_noon	The tweet is created from 11.am-13.pm	Boolean
	14. Is_posted_at_eve	The tweet is created from 6.pm-9.pm	Boolean
	15. Is_post_at_wee	The tweet is created at weekend	Boolean
Content-based	16. Contain_location	The tweet contains a location name	Boolean
	17. Contain_org	The tweet contains an organization name	Boolean
	18. Contain_tvshow	The tweet contains a television show name	Boolean
	19. Sentiment_level	The tweet is classified into sentiment levels	{positive, negative, objective}
	20. Contain_video	The tweet contains a video	Boolean
	21. Contain_picture	The tweet contains a picture	Boolean
	22. Contain_upper	The tweet contains upper words	Boolean
	23. Contain_number	The tweet contains number	Boolean
	24. Contain_excl	The tweet contains an exclamation mark	Boolean
	25. Contain_rt_term	The tweet contains RT term	Boolean
	26. Con_user_mentioned	The tweet mentions a user name	Boolean
	27. Contain_rt_sugges	The tweet contains one of the retweet suggestion term:Pls RT, please retweet, RT for..	Boolean
	28. Contain_URL[+]	The tweet contains an URL	Boolean
	29. Contain_hashtag[+]	The tweet contains a hashtag	Boolean
	30. Contain_famous_person	The tweet mentions a person who has been introduced on Wikipedia	Boolean
	31. Opt_length	Length of the content is between 70 to 100 characters	Boolean
	32. Len_of_text	Length of the content	#Numeric

With regard to user-based features, we consider that a Twitter user who highly interacts with other users will receive in turn high attention. We consider

the interaction between the user who sends the tweet and social networks through several features. First we use features Suh *et al.* defined [4]; they are marked by a$^+$ Table 1. We considered new several features that also aim at representing interaction between users. All the user-based features are numerical, although some are boolean values.

With regard to time-based features, we hypothesize that a majority of retweets are written shortly after the original tweet is posted and thus the posting time of a tweet may affect retweetability. Tweets that are posted in 'free hours' are more likely to receive more retweets than when posted in the official working hours. We define 4 time-base features considering the time that the tweet is posted. Each of these checks corresponds to a boolean feature in the tweet representation.

Finally, regarding content-based features, we defined 15 new content-based features. We considered named entity, sentiment level, media attachment, content enhancement, content size of the message. We also reused two existing features from Suh[4] which are marked by a$^+$ in the Table 1.

- *Named entity:* According to authors in [12], a location name mentioned in a tweet content could make the tweet more attractive. This may make people retweet the message. In addition, when a tweet mentions a famous person, it is likely to be retweeted by number of fans. We used Ritter's named entity extraction tool [13] to check whether the tweet contains a location name, an organization name, or a TV show reference. We also used this tool to recognize a person name in the tweet and then checked if this name is introduced as a person on DBpedia by an endpoint framework (http://dbpedia.org/snorql/). We hypothesize that well-known named entities contained in a tweet will make it more attractive and will be shared more. These features are boolean values.

- *Sentiment level:* We hypothesize tweets that are extremely positive or negative get the attention of people. We defined a new feature to capture the sentiment of tweets that we called Sentiment_level. We used a "scikit-learn" machine learning library (http://scikit-learn.org/stable/) to classify tweets as positive, negative or neutral.

- *Media attachment:* Twitter users often enrich their tweets by attaching media sources such as picture and video. We consider two boolean features to check whether the tweet contains a picture or a video.

- *Content enhancement:* There are some features that make the tweet stand out. We check whether the tweet contains an upper word, a number, or an exclamation mark. We also check if a tweet contains the 'RT' term, mentions a username, or includes words asking people to retweet a message such as 'please retweet', 'pls rt','retweet if','rt if','retweet to','rt to','rt!','retweet for','rt for', 'retweet'.

Finally, we also reuse two boolean features from [4] which check if the tweet contains a URL or a hashtag.

- *Content size:* The length of a tweet is limited to 140 character. We add a numeric feature to store the length of the tweet. We suppose that the tweet should not be too long or too short. There should be sufficient space for people

to add comments to the message when retweeting it. We check if the tweet has a suggested ideal length of 70 to 100 characters.

3.2 Machine learning model

We used different machine learning algorithms such as Naive Baiyes, Support Vector Machine, Decision Tree, and Random Forest (RF) implemented on Java Weka library (http://weka.sourceforge.net) In this paper, we report RF results only since they correspond to the best results we obtained, both for the baseline and for our model. For each collection, we used 10-folds cross validation.

4 Data and evaluation framework

We conducted experiments and evaluated our model on two types of collections: 1) 2 tweet collections about brand stories generated by consumers (namely 'Iphone' and 'Gucci' datasets) and 2)one tweet collection generated by the company who owns the brand ('Samsung' dataset). The first two datasets were extracted from 1-percent tweets dataset we collected from 21 September 2015 to 31 May 2017 using the keyword 'iphone' and 'gucci'. The last dataset was directly collected from official Twitter account '@SamsungMobileUS' using the keyword 'galaxy' from 21 Sept. 2015 to 9 Oct. 2017 through Twitter API.

Table 2. The number of tweets and their distribution for our three datasets used to evaluate our predictive model.

Table 3. Classes distribution of Iphone, Gucci and Samsung datasets used for multi-class classification.

	Iphone	Gucci	Samsung
# of tweets	2,188,923	242,956	19,231
# of non-retweeted tweets	1,483,705	74,543	14,311
# of (unique) retweeted tweets	312,003	51,805	4,920

	Iphone	Gucci	Samsung
Class-0	1,483,705	74,543	14,311
Class-1	271,147	41,752	4,625
Class-2	37,355	9,968	295
Class-3	501	85	0

Each tweet in these datasets is composed of several pieces of information regarding a twitter status such as unique identifier, text, time of posting, media and others. We used the value of the 'retweet_count' field which specifies the numbers of times a tweet has been retweeted to classify tweets in the predictive model (Section 5). Table 2 reports the number of tweets and their distribution for the three datasets.

The baseline model we report uses random forest on all Suh's features [4]. We compare it with the model that considers all the features we presented in Table 1 (our features plus Suh's features).

5 Experiments and Results

5.1 Binary classification

To predict if a given tweet will be retweeted or not, we classified tweets into two classes: class-0 corresponds to tweets that are not retweeted while class-1

are retweeted tweets. Since there is a huge difference between the number of tweets in the two classes (see Table 2), we balanced these numbers during the classification process.

For the Iphone and Samsung datasets, we divided each dataset into several sub-sets. The tweets from class-1 were all kept for all sub-sets while the tweets from class-0 were divided into sub-sets so that the number of tweets from class-0 is approximately the same as the number of tweets from class-1. For the Gucci dataset, since the number of tweets in class-0 is about one and a half the number of tweets from class-1, we generated synthetic samples from class-1 50%. More specifically, the balanced of classes is dealt as follows:

- *Iphone dataset.* The tweets from class-0 were divided into five sub-sets. Each sub-set includes the entire class-1 tweets (312,003 tweets) and one part class-0 tweets (about 296,741 tweets).

- *Gucci dataset.* We generated synthetic samples by randomly sampling attributes from instances from class-1 using Synthetic Minority Over-sampling Technique (SMOTE) on Weka. The settings for SMOTE are setNearestNeighbors = 5 and setPercentage = 50. As a result, the tweets from class-1 are one and a haft the original with 77,707 tweets from class-1 and 74,543 from class-0.

- *Samsung dataset.* The tweets from class-0 were divided into three parts. Each sub-set includes the entire class-1 tweets (4,920 tweets) and one part class-0 tweets (about 4,771 tweets). We had thus three sub-sets.

Table 4 reports the F-measure of the binary classification (a tweet is predicted to be retweeted or not) for the Iphone, Gucci and Samsung datasets. For the Iphone and Samsung datasets, we report the average of F-measure over the sub-sets (See Table 4).

Table 4. F-measure of the binary classification using Random Forest for three datasets. * indicates statistically significant differences by t-test compared to the baseline.

	Iphone			Gucci			Samsung		
	Cl-0	Cl-1	Aver.	Cl-0	Cl-1	Aver.	Cl-0	Cl-1	Aver.
Baseline	0.824	0.820	0.822	0.788	0.779	0.783	0.820	0.789	0.804
Our model (RF)	0.853	0.851	**0.852***	0.825	0.817	**0.821**	0.848	0.834	**0.841***

As it can be seen in the Table 4, we significantly improve the F-measure of the binary classification on average and on every class compared to the baseline for all datasets. Interestingly, both our model and baseline achieve higher performance on class-0 (tweets are not retweeted) than on class-1 (tweets are retweeted) even if the number of tweets from class-0 is smaller than those from class-1. Our method improves the results on class-1 more than on class-0 for the three datasets.

5.2 Multi-class classification

To predict the volume of retweets that a given tweet will receive in the future, we divided the tweets into four different classes like Hong *et al.* [8] and Hoang [14] did : class-0 (tweets that are not retweeted); class-1 (tweets that are retweeted

less than 100 times; class-2 (tweets that are retweeted less than 10,000 time and class-3 (tweets that are retweeted more than 10,000 times).

Table 3 presents the class distribution of the Iphone, Gucci and Samsung datasets. Similarly to the case of binary classification, number of tweets in classes are very imbalanced (see Table 3). We dealt with this problem as follow:

For the Iphone and Samsung datasets, we first divided each dataset into several sub-sets like we did with binary classification. The tweets from class-1, class-2 and class-3 (if any) were all kept for all sub-sets while the tweets from class-0 were divided into sub-sets so that the number of tweets from class-0 was approximately equal to those from class-1. Then, we SMOTE tweets from class-2 and class-3 100% (setNearestNeighbors = 5 and setPercentage = 100).

For the Gucci dataset, since the number of tweets from class-0 are about one and half the number of tweets from class-1, we SMOTE tweets from class-1 50% (setNearestNeighbors = 5, setPercentage = 50) and SMOTE tweets from class-2 and tweets from class-3 100% (setNearestNeighbors = 5, setPercentage = 100).

Table 5. F-measure of the multi-class classification using Random Forest for three datasets. * indicates statistically significant difference by t-test compared to the baseline.

Dataset	Class	Baseline	Our Method (RF)
Iphone	Cl0	0.821	0.849
	Cl1	0.719	0.761
	Cl2	0.588	0.640
	Cl3	0.130	0.114
	Av.	0.749	**0.787***
Gucci	Cl0	0.785	0.821
	Cl1	0.645	0.687
	Cl2	0.617	0.628
	Cl3	0.021	0.056
	Av.	0.707	**0.743***
Samsung	Cl0	0.848	0.847
	Cl1	0.774	0.793
	Cl2	0.513	0.731
	Cl3	–	–
	Av.	0.794	**0.816***

Table 5 reports the results of multi-class classification on the three datasets in terms of averaged F-measure over sub-sets.

Similarly to binary classification, our method highly improves the F-measure of the multi-class classification on average and on every class compared to the baseline for the three datasets. On average, comparing to the baseline, our method improves the F-measure from 2,2% to 3,8%, all statistically significant.

On each class of the three datasets, our method improves the F-measure compared to the baseline but with different effectiveness. We achieve high F-measure on class-0, class-1 and class-2 (from 0.628 to 0.847) but lower F-measure on class-3 (0.056 to 0.114) for the three datasets. This may be caused by the very huge difference of the number of tweets per class although we tried to limit this differences during training. In the Iphone and Gucci datasets, the number of

tweets from class-1 is about from four to seven times the number of tweets from class-2 and more than about five hundred times the number of tweets from class-3. In the Samsung dataset, the number of tweets from class-1 is about fifteen times the number of tweets from class-2 and there is no any tweets from class-3.

5.3 Most important features

Our predictive model uses 32 features of which we have proposed 25 features in this paper plus Suh's ones. We evaluated the importance of each feature by applying the Inforgain attribute evaluator using Ranker search method in Weka. This method calculates the relative weight of each feature in the model. The results are presented in the next subsections.

Binary classification. The best seven features when classifying tweets in binary classes are as follows. Numbers in brackets corresponds to the weight; the higher the value, the more important the feature is for the model.

- *Iphone dataset:* No_of_followers$^+$ (0.298), No_of_favourite$^+$ (0.116), No_of_followees$^+$ (0.093), Aver_favour_per_day (0.091), No_groups_user_belongs (0.084), Aver_tweets_per_day (0.066), Age_of_account$^+$ (0.062).

- *Gucci dataset:* No_of_followers$^+$ (0.242), No_groups_user-_belongs (0.168), Len_of_text (0.168), User_name_len (0.137), Aver_tweets_per_day (0.112), No_of_favourite$^+$ (0.108), Aver_-favour_per_day (0.089).

- *Samsung dataset:* No_of_followers$^+$ (0.607), Age_of_account$^+$ (0.545), Aver_favour_per_day (0.533), Aver_tweets_per_day (0.508), No_of_followees$^+$ (0.441), No_groups_user-_belongs (0.427), No_of_favourite$^+$ (0.328).

We found that one feature we reapply from Suh *et al.* (namely No_of_followers$^+$) is consistently the best feature on the three datasets. This result matches with their finding that the number of followers has a very strong relationship with the retweetability. Besides, the number of followees (No_of_followees$^+$) and age of account (Age_of_account$^+$), which are considered to be important in affecting to retweet rate by Suh, are also important features on the three datasets.

The number of tweets that the user posted in the past (Total_of_tweets$^+$) has not much impact on retweetability on both Suh finding and on ours. However, the number of tweets that the user has favourited in his time line was found to have very little impact on the retweet number by Suh *et al.* [4] while it is one of the best seven features on our three datasets. It could be interesting in future work to analyze the impact of the domain on this result.

One important result is that some of the new features we defined, number of groups or communities the user belongs to (No_groups_user_belongs), average tweets (Aver_tweets_per_day) and average likes the user makes a day (Aver_favour_per_day) are among the best 7 features whatever the dataset is.

The best features for the Iphone dataset are similar to those for the Samsung dataset with different weight. The situation is a little different in Gucci dataset. The length of text (Len_of_text) and user name (User_name_len) are important in Gucci dataset but not in the two other datasets. The reason might be that these features vary more in Gucci dataset than in the two other datasets.

Apart from the above features, the next important features on the three datasets with different weight are: User_is_verified, Total_of_tweets$^+$, Contain_hashtag $^+$, Contain_video, Contain_picture, Sentiment_level, Contain_upper.

Multi-class classification. Similarly to binary classification, two features from the literature No_of_followers$^+$, No_of_favourite$^+$ and one feature we defined (No_groups_user_belongs) are consistently in the best seven features.

More precisely, the best seven features when classifying tweets in multi-class classification are as follow:
- *Iphone dataset:* No_of_followers$^+$ (0.3414), Len_of_text (0.217), No_groups_user_belongs (0.199), No_of_favourite$^+$ (0.1504), User_name_len (0.1503),Aver_favour_per_day (0.142), No_of_followees$^+$ (0.137)
- *Gucci dataset:* No_of_followers$^+$ (0.316), No_groups_user-_belongs (0.215), Len_of_text (0.210), User_name_len (0.160), No_of_favourite$^+$ (0.125), Aver_favour_per_day (0.121), No_of_followees$^+$ (0.113)
- *Samsung dataset:* No_of_followers$^+$ (0.638), Age_of_account$^+$ (0.588), Aver_favour_per_day (0.571), Aver_tweets_per_day (0.546), No_groups_user-_belongs (0.478), No_of_followees$^+$ (0.452), No_of_favourite$^+$ (0.351).

As can be seen, the strong relationship between retweetability and No_of_followees$^+$ found by Suh's is confirmed again.

When considering the Iphone and Gucci datasets, the seven most important features in multi-class classification are similar; but relatively different from those for binary classification. For the Iphone dataset, length of text (Len_of_text) is not so important in the binary classification, but it is, in the multi-class classification while age of account (Age_of_account$^+$) which is one of the top 7 features in the binary classification is not so important in multi-classification. For the Gucci dataset, average number of tweets that the user posted per day (Aver_tweets_per_day) is fourth important in the binary classification but it has not much relationship with retweet rate in the multi-class classification; the number of followees (No_of_followees$^+$) is in the top seven features that impact on the possibility of retweet. Not surprisingly, the best features in the Samsung dataset are also important either in the Iphone or Gucci dataset; however two important features in these two datasets (Len_of_text and User_name_len) are not in the best seven features of Samsung dataset.

Apart from the above features, the next important features on three datasets with different weight are: User_is_verified, Total_of_tweets$^+$, Contain_hashtag $^+$, Contain_upper, Contain_video, Contain_picture, Sentiment_level.

5.4 Correlations between features

To evaluate if the new features we defined are dependent from existing features and independent from each others, we calculated the correlations between features. We applied the Principle Component evaluator using Ranker search method implemented on Weka. We obtained a correlation matrix which measures the degree of association between features for each dataset. The results are very similar from one data set to the others.

The first important point is that there are a few correlations that are significant; most of them are weak correlations. Most of the features are independent from each other. Indeed, most of the correlation values are between -0.2 to 0.2 for the three datasets.

There are three significant correlations across the three datasets, which are between features we defined and features from literature: No_groups_user_belongs correlates with No_of_followers$^+$, Aver_tweets_per_day correlates with Total_of_tweets$^+$ and User_is_verified correlates with No_of_followers$^+$. Except the first correlation, which includes two features that are important in the model; the two other correlations include features that are not significant in the model.

The other correlations are between Aver_tweets_per_day and Age_of_account$^+$ for the Samsung dataset and between Aver_favour_per_day and No_of_favourite$^+$ for the Gucci dataset. All the features in these correlations are important in the predictive model. Apart from that, the other significant correlations are among our features or between existing features and some features that we defined but that have little weights in the predictive model.

Some of the features that we developed in this paper are both significant for the predictive models (main features) and do not correlate with existing features from the literature. This is the case for Len_of_text and User_name_len.

6   Conclusions

This paper proposes a method that helps business managers to understand and predict how popular a message is in social networks. Specifically, we address the problem of predicting whether a given tweet will be retweeted or not, and the challenge of predicting the volume of retweets that a certain tweet will receive.

We defined and developed new features in addition to the ones from the literature and applied a machine learning model using several classifiers. Our features are grouped into three types: user-based, time-based and content-based features. Using two types of collections: consumer-generate stories and company's official stories, we show that our model significantly improves and by about 4% the F-measure compared to the state of art for both binary classification and multi-class classification when evaluated on two types of collections: collections of brand stories generated by consumers and by the owner of the brand.

There are some features that are more important than others. We show that the number of followers, followees, favourites of the user and the number of groups that the user belongs to, are the most important features in making a tweet about a brand story to be retweeted. In addition, length of message, containing hashtag, URL, famous person and picture also correlate with the retweetability. We recommend to combine these features to make a message widely spread in social networks.

Indeed, we also analyzed the correlations between features in the three datasets. Most of features are independent from each others. The few features of ours that correlate with existing features, have generally low weights when analyzing their impact for the predictive models. In addition, the results presented in section 5

show that the combination of the features we defined and existing features significantly improves the performance of the predictive model.

We believe that, our model will help business managers to predict the diffusion of information related to their brand/products in social networks. In addition, this paper also proposes features that make a message popular. This would help business managers to form stories online to broadcast their brand/products as well as propose strategies to control or promote customer-generated stories.

References

1. W. Assaad and J. M. Gomez, "Social network in marketing (social media marketing) opportunities and risks," *International Journal of Managing Public Sector Information and Communication Technologies*, vol. 2, no. 1, p. 13, 2011.
2. W. G. Mangold and D. J. Faulds, "Social media: The new hybrid element of the promotion mix," *Business horizons*, vol. 52, no. 4, pp. 357–365, 2009.
3. S. Gensler, F. Völckner, Y. Liu-Thompkins, and C. Wiertz, "Managing brands in the social media environment," *J. of Interactive Marketing*, vol. 27, no. 4, pp. 242–256, 2013.
4. B. Suh, L. Hong, P. Pirolli, and E. H. Chi, "Want to be retweeted? large scale analytics on factors impacting retweet in twitter network," in *Social computing (socialcom), 2010 ieee second international conference on*, pp. 177–184, IEEE, 2010.
5. Y. Hu, C. Hu, S. Fu, P. Shi, and B. Ning, "Predicting the popularity of viral topics based on time series forecasting," *Neurocomputing*, vol. 210, pp. 55–65, 2016.
6. H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?," in *Proceedings of the 19th international conference on World wide web*, pp. 591–600, ACM, 2010.
7. C. Remy, N. Pervin, F. Toriumi, and H. Takeda, "Information diffusion on twitter: everyone has its chance, but all chances are not equal," in *Signal-Image Technology & Internet-Based Systems (SITIS)*, pp. 483–490, IEEE, 2013.
8. L. Hong, O. Dan, and B. D. Davison, "Predicting popular messages in twitter," in *Int. Conf. Companion on WWW*, pp. 57–58, ACM, 2011.
9. X. Ren and Y. Zhang, "Predicting information diffusion in social networks with users social roles and topic interests," in *Information Retrieval Technology*, pp. 349–355, Springer, 2016.
10. J. Zhang, B. Liu, J. Tang, T. Chen, and J. Li, "Social influence locality for modeling retweeting behaviors.," in *IJCAI*, vol. 13, pp. 2761–2767, 2013.
11. Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, and Z. Su, "Understanding retweeting behaviors in social networks," in *Int. conf. on Information and knowledge management*, pp. 1633–1636, ACM, 2010.
12. J. Lingad, S. Karimi, and J. Yin, "Location extraction from disaster-related microblogs," in *Int. Conf. on WWW*, pp. 1017–1020, ACM, 2013.
13. A. Ritter, S. Clark, O. Etzioni, *et al.*, "Named entity recognition in tweets: an experimental study," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1524–1534, ACL, 2011.
14. T. B. N. Hoang and J. Mothe, "Predicting Information Diffusion on Twitter - Analysis of predictive features," *Journal of Computational Science*, vol. 22, 2017.